NL020356

Europäisches Patentamt | European Patent Office | Office européen des brevets

IB/03/1650

Bescheinigung | Certificate | Attestation

PRIORITY DOCUMENT 21 JUL 2003
WIPO PCT

Die angehefteten Unterlagen stimmen mit der ursprünglich eingereichten Fassung der auf dem nächsten Blatt bezeichneten europäischen Patentanmeldung überein.

The attached documents are exact copies of the European patent application described on the following page, as originally filed.

Les documents fixés à cette attestation sont conformes à la version initialement déposée de la demande de brevet européen spécifiée à la page suivante.

Patentanmeldung Nr. Patent application No. Demande de brevet n°

02076588.9

PRIORITY DOCUMENT
SUBMITTED OR TRANSMITTED IN COMPLIANCE WITH RULE 17.1(a) OR (b)

BEST AVAILABLE COPY

Der Präsident des Europäischen Patentamts;
Im Auftrag

For the President of the European Patent Office

Le Président de l'Office européen des brevets
p.o.

R C van Dijk

Anmeldung Nr:
Application no.: 02076588.9
Demande no:

Anmeldetag:
Date of filing: 22.04.02
Date de dépôt:

Anmelder/Applicant(s)/Demandeur(s):

Koninklijke Philips Electronics N.V.
Groenewoudseweg 1
5621 BA Eindhoven
PAYS-BAS

Bezeichnung der Erfindung/Title of the invention/Titre de l'invention:
(Falls die Bezeichnung der Erfindung nicht angegeben ist, siehe Beschreibung.
If no title is shown please refer to the description.
Si aucun titre n'est indiqué se referer à la description.)

Spatial audio

In Anspruch genommene Priorität(en) / Priority(ies) claimed /Priorité(s) revendiquée(s)
Staat/Tag/Aktenzeichen/State/Date/File no./Pays/Date/Numéro de dépôt:

Internationale Patentklassifikation/International Patent Classification/
Classification internationale des brevets:

H04S/

Am Anmeldetag benannte Vertragstaaten/Contracting states designated at date of filing/Etats contractants désignées lors du dépôt:

AT BE CH CY DE DK ES FI FR GB GR IE IT LI LU MC NL PT SE TR

Spatial audio

Background

Prior solutions in audio coders that have been suggested to reduce the bitrate of stereo program material include:

'Intensity stereo'. In this algorithm, high frequencies (typically above 5 kHz) are represented by a single audio signal (i.e., mono), combined with time-varying and frequency-dependent scalefactors.

'M/S stereo'. In this algorithm, the signal is decomposed into a sum (or mid, or common) and a difference (or side, or uncommon) signal. This decomposition is sometimes combined with principle component analysis or time-varying scalefactors. These signals are then coded independently, either by a transform coder or waveform coder. The amount of information reduction achieved by this algorithm strongly depends on the spatial properties of the source signal. For example, if the source signal is monaural, the difference signal is zero and can be discarded. However, if the correlation of the left and right audio signals is low (which is often the case), this scheme offers only little advantage.

Parametric descriptions of audio signals have gained interest during the last years, especially in the field of audio coding. It has been shown that transmitting (quantized) parameters that describe audio signals requires only little transmission capacity to resynthesize a perceptually equal signal at the receiving end. However, current parametric audio coders focus on coding monaural signals, and stereo signals are often processed as dual mono.

Invention

According to an aspect of the invention, *spatial attributes* of multichannel audio signals are parameterized. It will be shown that for general audio coding applications, transmitting these parameters combined with only *one* monaural audio signal will strongly reduce the transmission capacity necessary to transmit the stereo signal compared to audio coders that process the channels independently, while maintaining the original spatial impression. An important issue is that although people receive waveforms of an auditory object twice (once by the left ear and once by the right ear), only a single auditory object is

perceived at a certain position and with a certain size (or spatial diffuseness). Therefore, it seems unnecessary to describe audio signals as two or more (independent) waveforms and it would be better to describe multichannel audio as a set of auditory objects, each with its own spatial properties. One difficulty that immediately arises is the fact that it is almost impossible to automatically separate individual auditory objects from a given ensemble of auditory objects, for example a musical recording. This problem can be circumvented by not splitting the program material in individual auditory objects, but rather describing the spatial parameters in a way that resembles the effective (peripheral) processing of the auditory system. In particular, the parametric description of multichannel audio presented here is related to the binaural processing model presented by Breebaart et al. This model aims at describing the effective signal processing of the binaural auditory system. For a complete model description, see Breebaart et al (2001a,b,c). A short interpretation is given below which helps to understand the invention.

The model splits the incoming audio into several band-limited signals, which are (preferably) spaced linearly at an ERB-rate scale. The bandwidth of these signals depends on the center frequency, following the ERB rate. Subsequently, preferably *for every frequency band*, the following properties of the incoming signals are analyzed:

- The interaural level difference, or ILD, defined by the relative levels of the band-limited signal stemming from the left and right ears,
- The interaural time (or phase) difference (ITD or IPD), defined by the interaural delay (or phase shift) corresponding to the peak in the interaural cross-correlation function, and
- The (dis)similarity of the waveforms that can not be accounted for by ITDs or ILDs, which can be parameterized by the *maximum* interaural cross-correlation (i.e., the value of the cross-correlation at the position of the maximum peak).

The three parameters described above vary over time; however, since the binaural auditory system is very sluggish in its processing, the update rate of these properties is rather low (typically tens of milliseconds).

It may be assumed here that the (slowly) time-varying properties mentioned above are the *only* spatial signal properties that the binaural auditory system has available, and that from these time and frequency dependent parameters, the perceived auditory world is reconstructed by higher levels of the auditory system.

It is interesting to mention that the ILD and ITD are believed to be the most important localization cues in the horizontal plane, while the maximum interaural cross-

correlation is strongly related to the *perceptual spatial diffuseness* (or compactness) of a sound source.

It is an insight of the inventors that it is sufficient to describe spatial attributes of any multichannel audio signal by specifying the ILD, ITD (or IPD) and maximum correlation *as a function of time and frequency*.

An embodiment of the current invention aims at describing a multichannel audio signal by:

one monaural signal, consisting of a certain combination of the input signals, and

a set of spatial parameters: two localization cues (ILD, and ITD or IPD) and a parameter that describes the similarity or dissimilarity of the waveforms that cannot be accounted for by ILDs and/or ITDs (e.g., the maximum of the cross-correlation function) preferably for every time/frequency slot. Preferably, spatial parameters are included for each additional auditory channel.

Advantages of this parametric description are the following:

- Decoupling of monaural and binaural signal parameters in audio coders. Difficulties related to stereo audio coders are strongly reduced (such as the audibility of interaurally uncorrelated quantization noise compared to interaurally correlated quantization noise).
- Strong bitrate reduction in audio coders due to a low update rate and low frequency resolution required for the spatial parameters. The associated bitrate to code the spatial parameters is typically 10 kbit/s or less (see embodiment).
- Easy combination with existing audio coders. The proposed scheme produces one mono signal that can be coded and decoded with any existing coding strategy. After monaural decoding, the system described here regenerates the spatial attributes.

The set of spatial parameters can be used as an enhancement layer in audio coders. For example, a mono signal is transmitted if only a low bitrate is allowed, while by including the spatial enhancement layer the decoder can reproduce stereo sound.

The invention can in principle be used to generate n channels from one mono signal, if $(n-1)$ sets of spatial parameters are transmitted. In such condition, the spatial parameters describe how to form the n different audio channels from the single mono signal.

Analysis methods

In the following, it is assumed that the incoming signals are split up in band-pass signals (preferably with a bandwidth which increases with frequency) and that

parameters can be analyzed as a function of time. A possible method for time/frequency slicing would be to use time-windowing followed by a transform operation, but also time-continuous methods could be used (e.g., filterbanks). The next steps consist of (1) finding the level difference (ILD) of corresponding subband signals, (2) finding the time difference (ITD or IPD) of corresponding subband signals, and (3) describe the amount of similarity or dissimilarity of the waveforms which cannot be accounted for by ILDs or ITDs. The analysis of these parameters is discussed below.

Analysis of ILDs

The ILD is determined by the level difference of the signals at a certain time instance for a given frequency band. One method to determine the ILD is to measure the rms value of the corresponding frequency band of both input channels and compute the ratio of these rms values (preferably expressed in dB).

Analysis of the ITDs

The ITDs are determined by the time or phase alignment which gives the best match between the waveforms of both channels. One method to obtain the ITD is to compute the cross-correlation function between two corresponding subband signals and searching for the maximum. The delay that corresponds to this maximum in the cross-correlation function can be used as ITD value. A second method would be to compute the analytic signals of the left and right subband (i.e., computing phase and envelope values) and use the phase difference between the channels as IPD parameter.

Analysis of the correlation

The correlation is obtained by first finding the ILD and ITD that gives the best match between the corresponding subband signals and subsequently measuring the similarity of the waveforms after compensation for the ITD and/or ILD. Thus, in this framework, the correlation is defined as the *similarity or dissimilarity of corresponding subband signals which can not be attributed to ILDs and/or ITDs*. A suitable measure for this parameter is the maximum value of the cross-correlation function (i.e., the maximum across a set of delays). However, also other measures could be used, such as the relative energy of the difference signal after ILD and/or ITD compensation compared to the sum signal of corresponding subbands (preferably also compensated for ILDs and/or ITDs). This difference parameter is basically a linear transformation of the (maximum) correlation.

Parameter quantization

An important issue of transmission of parameters is the accuracy of the parameter representation (i.e., the size of quantization errors), which is directly related to the necessary transmission capacity. In this section, several issues with respect to the quantization of the spatial parameters will be discussed. The basic idea is to base the quantization errors on so-called *just-noticable differences* (JNDs) of the spatial cues. To be more specific, the quantization error is determined by the sensitivity of the human auditory system to changes in the parameters. Since it is well known that the sensitivity to changes in the parameters strongly depends on the values of the parameters itself, we apply the following methods to determine the discrete quantization steps.

Quantization of ILDs

It is known from psychoacoustic research that the sensitivity to changes in the IID depends on the ILD itself. If the ILD is expressed in dB, deviations of approximately 1 dB from a reference of 0 dB are detectable, while changes in the order of 3 dB are required if the reference level difference amounts 20 dB. Therefore, *quantization errors can be larger if the signals of the left and right channels have a larger level difference*. For example, this can be applied by first measuring the level difference between the channels, followed by a non-linear (compressive) transformation of the obtained level difference and subsequently a linear quantization process, or by using a lookup table for the available ILD values which have a nonlinear distribution. The embodiment below gives an example of such a lookup table.

Quantization of the correlation

The quantization error of the correlation depends on (1) the correlation value itself and possibly (2) on the ILD. Correlation values near +1 are coded with a high accuracy (i.e., a small quantization step), while correlation values near 0 are coded with a low accuracy (a large quantization step). An example of a set of non-linearly distributed correlation values is given in the embodiment. A second possibility is to use quantization steps for the correlation that depend on the measured ILD *of the same subband*: for large ILDs (i.e., one channel is dominant in terms of energy), the quantization errors in the correlation become larger. An extreme example of this principle would be to not transmit correlation values for a certain subband at all if the absolute value of the IID for that subband is beyond a certain threshold.

Quantization of the ITDs

The sensitivity to changes in the ITDs of human subjects can be characterized as having a constant phase threshold. This means that in terms of delay times, the quantization steps for the ITD should decrease with frequency. Alternatively, if the ITD is represented in the form of phase differences, the quantization steps should be independent of frequency. One method to implement this would be to take a fixed phase difference as quantization step and determine the corresponding time delay for each frequency band. This ITD value is then used as quantization step. Another method would be to transmit phase differences which follow a frequency-independent quantization scheme. It is also known that above a certain frequency, the human auditory system is not sensitive to ITDs in the finestructure waveforms. This phenomenon can be exploited by only transmitting ITD parameters up to a certain frequency (typically 2 kHz).

A third method of bitstream reduction is to incorporate ITD quantization steps that depend on the ILD and /or the correlation parameters of the same subband. For large ILDs, the ITDs can be coded less accurately. Furthermore, if the correlation it very low, it is known that the human sensitivity to changes in the ITD is reduced. Hence larger ITD quantization errors may be applied if the correlation is small. An extreme example of this idea is to not transmit ITDs at all if the correlation is below a certain threshold.

Embodiment

The embodiment for a stereo input signal can be schematically drawn as shown in Fig. 1.

Fig. 1. Schematic diagram of an embodiment of the invention. In the encoder, spatial parameters are analyzed preferably for each time/frequency slot. Subsequently, a sum (or dominant) signal is generated consisting of a certain combination of the at least two input signals. Synthesis (decoder) is performed by applying the spatial parameters to the sum signal to generate left and right output signals.

In this embodiment, the spatial parameter description is combined with a monaural (single channel) audio coder to encode a stereo audio signal. It should be noted that although the described embodiment works on stereo signals, the general idea can be applied to n-channel audio signals, with $n>1$.

Analysis

The left and right incoming signals are split up in various time frames (2048 samples at 44.1 kHz sampling rate) and windowed with a square-root Hanning window. Subsequently, FFTs are computed. The negative FFT frequencies are discarded and the resulting FFTs are subdivided into groups (subbands) of FFT bins. The number of FFT bins that are combined in a subband g depends on the frequency: at higher frequencies more bins are combined than at lower frequencies. In the current implementation, FFT bins corresponding to approximately 1.8 ERBs (Equivalent Rectangular Bandwidth) are grouped, resulting in 20 subbands to represent the entire audible frequency range. The resulting number of FFT bins $S[g]$ of each subsequent subband (starting at the lowest frequency) is

S=[4 4 4 5 6 8 9 12 13 17 21 25 30 38 45 55 68 82 100 477]

Thus, the first three subbands contain 4 FFT bins, the fourth subband contains 5 FFT bins, etc. For each subband, the corresponding ILD, ITD and correlation (r) are computed. The ITD and correlation are computed simply by setting all FFT bins which belong to other groups to zero, multiplying the resulting (band-limited) FFTs from the left and right channels, followed by an inverse FFT transform. The resulting cross-correlation function is scanned for a peak within an interchannel delay between –64 and +63 samples. The internal delay corresponding to the peak is used as ITD value, and the value of the cross-correlation function at this peak is used as this subband's interaural correlation. Finally, the ILD is simply computed by taking the power ratio of the left and right channels for each subband.

Generation of the sum signal

The left and right subbands are summed after a phase correction (temporal alignment). This phase correction follows from the computed ITD for that subband and consists of delaying the left-channel subband with ITD/2 and the right-channel subband with –ITD/2. The delay is performed in the frequency domain by appropriate modification of the phase angles of each FFT bin. Subsequently, the sum signal is computed by adding the phase-modified versions of the left and right subband signals. Finally, to compensate for uncorrelated or correlated addition, each subband of the sum signal is multiplied with sqrt(2/(1+r)), with r the correlation of the corresponding subband. If necessary, the sum signal can be converted to the time domain by (1) inserting complex conjugates at negative frequencies, (2) inverse FFT, (3) windowing, and (4) overlap-add.

Quantization of spatial parameters

ILDs (in dB) are quantized to the closest value out of the following set I:

I=[-19 -16 -13 -10 -8 -6 -4 -2 0 2 4 6 8 10 13 16 19]

ITD quantization steps are determined by a constant phase difference in each subband of 0.1 rad. Thus, for each subband, the time difference that corresponds to 0.1 rad of the subband center frequency is used as quantization step. For frequencies above 2 kHz, no ITD information is transmitted.

Interaural correlation values r are quantized to the closest value of the following ensemble R:

R=[1 0.95 0.9 0.82 0.75 0.6 0.3 0]

This will cost another 3 bits per correlation value.

If the absolute value of the (quantized) ILD of the current subband amounts 19 dB, no ITD and correlation values are transmitted for this subband. If the (quantized) correlation value of a certain subband amounts zero, no ITD value is transmitted for that subband.

In this way, each frame requires a maximum of 233 bits to transmit the spatial parameters. With a framelength of 1024 frames, the maximum bitrate for transmission amounts 10.25 kbit/s. It should be noted that using entropy coding or differential coding, this bitrate can be reduced further.

Synthesis

In this part, it is assumed that the frequency-domain representation of the sum signal as described in the analysis section is available for processing. This representation may be obtained by windowing and FFT operations of the time-domain waveform. First, the sum signal is copied to the left and right output signals. Subsequently, the correlation between the left and right signals is modified with a decorrelator. Subsequently, each subband of the left signal is delayed by -ITD/2, and the right signal is delayed by ITD/2 given the (quantized) ITD corresponding to that subband. Finally, the left and right subbands are scaled according to the ILD for that subband. To convert the output signals to the time domain, the following steps have to be performed: (1) inserting complex conjugates at negative frequencies, (2) inverse FFT, (3) windowing, and (4) overlap-add.

In summary, this application describes a psycho-acoustically motivated, parametric description of the spatial attributes of multichannel audio signals. This parametric

description allows strong bitrate reductions in audio coders, since only one monaural signal has to be transmitted, combined with (quantized) parameters which describe the spatial properties of the signal. The decoder can form the original amount of audio channels by applying the spatial parameters. For near-CD-quality stereo audio, a bitrate associated with these spatial parameters of 10 kbit/s or less seems sufficient to reproduce the correct spatial impression at the receiving end.

References

Breebaart, J., van de Par, S. and Kohlrausch, A. (2001a). Binaural processing model based on contralateral inhibition. I. Model setup. *J. Acoust. Soc. Am.*, **110**, 1074-1088

Breebaart, J., van de Par, S. and Kohlrausch, A. (2001b). Binaural processing model based on contralateral inhibition. II. Dependence on spectral parameters. *J. Acoust. Soc. Am.*, **110**, 1089-1104

Breebaart, J., van de Par, S. and Kohlrausch, A. (2001c). Binaural processing model based on contralateral inhibition. III. Dependence on temporal parameters.. *J. Acoust. Soc. Am.*, **110**, 1105-1117

CLAIMS:

1. A method of coding an audio signal, the method comprising:

generating a monaural signal comprising a certain combination of at least two input audio channels,

analyzing spatial parameters of the at least two input audio channels, preferably for each time/frequency slot, to obtain a set of spatial parameters preferably for every time/frequency slot, the set including at least two localization cues (e.g. ILD, and ITD or IPD) and a parameter that describes a similarity or dissimilarity of waveforms that cannot be accounted for by the localization cues, the parameter being e.g. a maximum of a cross-correlation function, and

generating an encoded signal comprising the monoaural signal and the set of spatial parameters.

2. An encoder for coding an audio signal, the encoder comprising:

means for generating a monaural signal comprising a certain combination of at least two input audio channels,

means for analyzing spatial parameters of the at least two input audio channels, preferably for each time/frequency slot, to obtain a set of spatial parameters preferably for every time/frequency slot, the set including at least two localization cues (e.g. ILD, and ITD or IPD) and a parameter that describes a similarity or dissimilarity of waveforms that cannot be accounted for by the localization cues, the parameter being e.g. a maximum of a cross-correlation function, and

means for generating an encoded signal comprising the monoaural signal and the set of spatial parameters.

3. An apparatus for supplying an audio signal, the apparatus comprising:

an input for receiving an audio signal,

an encoder as claimed in claim 2 for encoding the audio signal to obtain an encoded audio signal, and

an output for supplying the encoded audio signal.

4. An encoded audio signal, the signal comprising:

a monaural signal comprising a certain combination of at least two audio channels, and

a set of spatial parameters, preferably for every time/frequency slot, the set including at least two localization cues (e.g. ILD, and ITD or IPD) and a parameter that describes a similarity or dissimilarity of waveforms that cannot be accounted for by the localization cues, the parameter being e.g. a maximum of a cross-correlation function.

5. A storage medium on which an encoded signal as claimed in claim 4 has been stored.

6. A method of decoding an encoded audio signal, the method comprising:

obtaining a monaural signal from the encoded audio signal, the monaural signal comprising a certain combination of at least two audio channels, and

obtaining a set of spatial parameters from the encoded audio signal, preferably for every time/frequency slot, the set including at least two localization cues (e.g. ILD, and ITD or IPD) and a parameter that describes a similarity or dissimilarity of waveforms that cannot be accounted for by the localization cues, the parameter being e.g. a maximum of a cross-correlation function, and

applying the spatial parameters to the monaural signal or the at least two audio channels to generate a multi-channel output signal.

7. A decoder for decoding an encoded audio signal

means for obtaining a monaural signal from the encoded audio signal, the monaural signal comprising a certain combination of at least two audio channels, and

means for obtaining a set of spatial parameters from the encoded audio signal, preferably for every time/frequency slot, the set including at least two localization cues (e.g. ILD, and ITD or IPD) and a parameter that describes a similarity or dissimilarity of waveforms that cannot be accounted for by the localization cues, the parameter being e.g. a maximum of a cross-correlation function, and

means for applying the spatial parameters to the monaural signal or the at least two audio channels to generate a multi-channel output signal.

8. An apparatus for supplying a decoded audio signal, the apparatus comprising:

an input for receiving an encoded audio signal,

a decoder as claimed in claim 7 for decoding the encoded audio signal to obtain a multi-channel output signal,

an output for supplying or reproducing the multi-channel output signal.

ABSTRACT:

In summary, this application describes a psycho-acoustically motivated, parametric description of the spatial attributes of multichannel audio signals. This parametric description allows strong bitrate reductions in audio coders, since only one monaural signal has to be transmitted, combined with (quantized) parameters which describe the spatial properties of the signal. The decoder can form the original amount of audio channels by applying the spatial parameters. For near-CD-quality stereo audio, a bitrate associated with these spatial parameters of 10 kbit/s or less seems sufficient to reproduce the correct spatial impression at the receiving end.

Sole Fig.

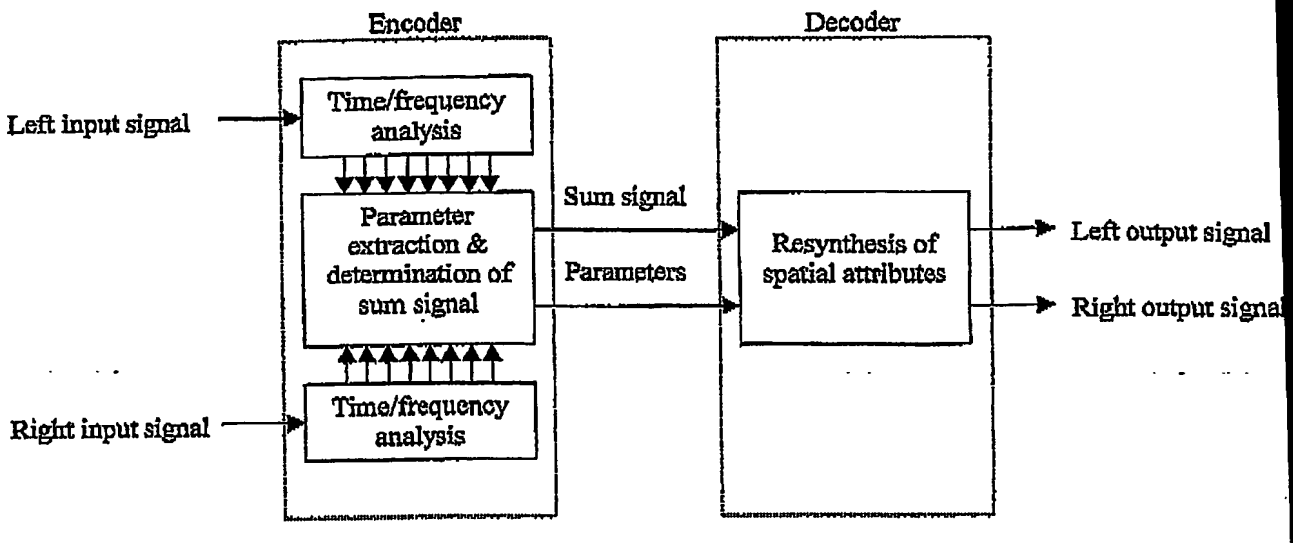
Encoder
Decoder
Left input signal
Time/frequency analysis
Parameter extraction & determination of sum signal
Time/frequency analysis
Right input signal
Sum signal
Parameters
Resynthesis of spatial attributes
Left output signal
Right output signal

Fig. 1